Developing Metrics and Evaluation Methods for Assessing AI-Enabled Robots in Manufacturing

Adam Norton, Amy Saretsky, and Holly Yanco
New England Robotics Validation and Experimentation (NERVE) Center
University of Massachusetts Lowell
110 Canal Street, Lowell, Massachusetts 01852
Corresponding author e-mail: adam_norton@uml.edu

Abstract

Evaluating the capabilities of a robotic system for manufacturing can include metrics related to performance, efficiency, and productivity. Measures for traditional industrial automation typically address operations that rely on strict repetition that does not allow for much variation. The inclusion of artificial intelligence (AI) in robotic systems can allow for greater aptitude in maintaining capability in the presence of variation, such as local changes in environmental characteristics or global changes in task execution parameters. New evaluation methods and metrics are needed to allow these advanced capabilities to be appropriately measured. This paper discusses evaluating the robustness, adaptability, generalizability, and versatility of AI-enabled robotic manufacturing systems. The considerations for conducting evaluations of these capabilities are reviewed, including implications for robots that learn and those that are designed to be explainable. Recommendations are made for advancing the development of metrics and evaluation methods that highlight the capabilities afforded by AI. A prototype framework is presented to guide the design of evaluations and classification of metrics.

Introduction

Traditional robot automation in manufacturing performs the same task over and over, allowing for highly repeatable metrics related to performance, efficiency, and productivity. Such systems may not be robust in the presence of variation (e.g., a target object is not in the exact place it is expected to be) or may not be able to be reconfigured for other tasks. The advent of robots with artificial intelligence (AI), or AI-enabled robots, in manufacturing enables agile and flexible solutions that can adapt to variation or uncertainty (El-Maraghy 2005; Browne et al. 1984). Variation can appear in many forms including fluctuations in environmental or task characteristics, which can be expected and trained for or unexpected and must be acclimated to.

Metrics and evaluation methods are needed to measure and express the capabilities of advanced robotics in manufacturing and to induce variation that appropriately demonstrates those capabilities. Robots in this domain may also be outfitted with learning capabilities to improve their performance and may be tasked with explaining their behavior. Both capabilities require special attention when designing evaluations. Prior work in test and evaluation can be leveraged from relevant domains including those for industrial manipulators, autonomous industrial vehicles, human-robot interaction (HRI), and machine learning.

This paper presents some of the considerations for developing metrics and evaluation methods for measuring robot capabilities that are enabled by AI, particularly those that operate in the presence of variation. These variations aid in defining the context of a manufacturing operation/evaluation and can include changes to:

- the input data provided to the robot to perform its task,
- the target objects being interacted with,
- the tasks being performed with those objects,
- the environment where the tasks are being executed, and
- the robot platform executing the tasks.

These variances must be properly characterized so that they accurately represent the context in which a robot will operate (Norton, Messina, and Yanco 2020 In Press). This process is paramount to eliciting results that are potentially generalizable to other, similar scenarios (Amigoni, Luperto, and Schiaffonati 2017), rather than abstract test cases. To do so, the parameters must be selected, measured, and induced as part of an evaluation.

This paper is primarily concerned with outcomes-based measures (i.e., those more directly observable) rather than internal assessments of AI. Considerations for developing metrics and evaluation methods are discussed, including implications if the robot possesses learning capabilities or is designed to be explainable. Several recommendations are made followed by a proposed framework for guiding the design of evaluations and classification of metrics.

Related Metrics and Evaluation Methods

Performance evaluation is critical to many robotics domains including mobile vehicle navigation, manipulation, and human-robot interaction. Some of the metrics and evaluation methods used in these domains are applicable to AI-enabled robots in manufacturing.

The capabilities measured for autonomous mobile vehicles include navigation, obstacle avoidance, and localization. Test methods such as those presented in (Norton, Gavriel, and Yanco 2019) vary the shape of the boundaries defining the environment in order to exercise these capabilities and account for dynamic changes like obstacle presence. Relevant metrics include the distance maintained from obstacles, dimensions of the space being navigated, robot trajectory smoothness, and traversal time (Ceballos, Valencia, and Ospina 2010); repeatability of the latter is particularly important for manufacturing environments (Gill et al. 2019).

Evaluating robotic manipulation capabilities includes measures related to grasping and functional task performance like assembly. A series of test methods for measuring the characteristics of robot end effectors including strength, cycle time, and repeatability are discussed in (Falco, Van Wyk, and Messina 2018; Falco et al. 2020a). Those three metrics in particular are very important for manufacturing; e.g., timing informs productivity throughput, repeatability influences expected error rates. Protocols have also been developed for pick and place or kitting operations, attempting to generalize a protocol that can be adapted for particular contexts (Mahler et al. 2018) and utilizing task-specific constraints to determine a success metric (Ortenzi et al. 2019). Common object sets for benchmarking manipulation (Calli et al. 2015) are also widely used in the community. Test methods and protocols for robotic assembly of small parts have also been developed (Falco et al. 2020b) which represent various manipulation and perception competencies for locating objects, inserting, nut threading, etc.; see Figure 1. The boards can be used for assembly and disassembly operations.

There are also efforts to develop test methods for mobile manipulators in manufacturing settings. A reconfigurable mobile manipulator artifact (RMMA) is specified in (Bostelman et al. 2016) that can be used to represent positioning requirements of assembly tasks. Performance is evaluated by motion capture and by measuring uncertainty of end effector position using reflectors. Test methods for agility have been developed to test a robot's ability to perform kitting tasks in the presence of induced variation (Downs, Harrison, and Schlenoff 2016). Metrics include awareness of parts having been dropped, correcting the task by picking up a part having been dropped, and acclimation to new prioritization of tasks (e.g., does the robot optimize for efficiency by fulfilling the new order using the kit it is already building?). More performance metrics for agility in smart manufacturing are presented in (Lee et al. 2017) which include metrics related to environmental characteristics, quality, flexibility, and adaptability.

Additional relevant fields for AI-enabled robots in manufacturing include metrics for HRI (Steinfeld et al. 2006), particularly those that could be applied to co-located scenarios in manufacturing (e.g., operator interventions to correct autonomous failures, the robot's awareness of the human and vice versa). Metrics for explainable AI are also relevant for robots that can learn in a manufacturing setting, such as measuring the "goodness" of an explanation, user satisfaction, understandability, trust, and human-robot task performance (Hoffman et al. 2018). The referenced paper reviews many tools that can be used for analysis including surveys for humans to rate explanations and a framework for conducting evaluations of explanations.

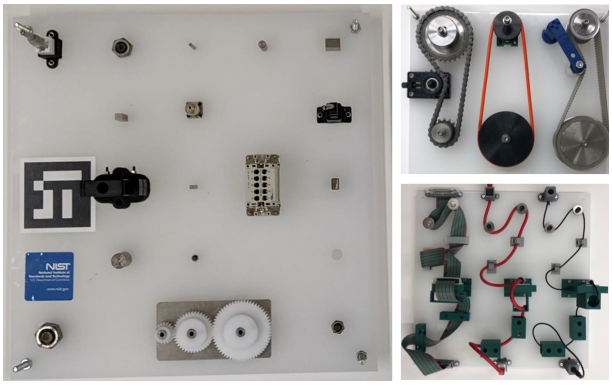

Figure 1: Assembly task board artifacts used to benchmark robotic assembly capabilities. Adapted from (Falco et al. 2020b) with permission.

Selecting Parameters

When selecting which parameters will be varied, it is important that those selected are (1) modeled after expectations of the real world scenario and (2) when adjusted from a baseline, the introduced variation is expected to challenge the system. For example, randomized bin picking has several parameters that are expected to affect robot capability, including scene complexity, degrees of freedom of the object's pose, image feature complexity, and part rigidity (Marvel et al. 2012). Each of the parameters can also be tied to robotic components whose capability can be exercised; e.g., scene complexity for vision, part rigidity for grasping. Manufacturing processes can be broken down and classified into the unique capabilities required to perform them to use as a starting point for selecting relevant parameters. Taxonomies of tasks like those used in (Shneier et al. 2015) for robotic assembly can be leveraged or similar exercises can be performed to derive the performance requirements of other domains. That paper draws from (Boothroyd, Dewhurst, and Knight 2010), which features many details of manufacturing processes including assembly, machining, injection molding, casting, and forging.

It is infeasible to test all possible variations of a parameter, even within a limited threshold, so a downselection of reasonably varied test cases must be performed. Given how thoroughly manufacturing processes are defined, it is likely that the expected amount of uncertainty could be modeled to generate various unique test cases. There remains a challenge to determine how many unique test cases should be selected and what they should consist of in order to prevent exhaustive evaluations while also sufficiently covering the possibilities. When varying a single parameter, one option is to utilize maximum, minimum, and median settings along that parameter, the former two representing edge cases. This method is utilized in (Falco et al. 2020a) to evaluate the

grasp strength of a robotic end effector by testing with artifacts that represent the largest, smallest, and average size object that can be grasped.

As more parameters get added, one could generate all possible combinations of settings for each parameter then eliminate combinations that are unrealistic, but this could still lead to exhaustive, expensive testing. The unique parameters of a target object or task according to their related application domain can influence the appropriate selection of parameters. For example, (Malhan et al. 2019) utilizes five target objects of various dimensions and shape that represent types of parts for composite layup in aerospace manufacturing (see Figure 2). Each object can be classified as requiring different types of trajectories to be planned across parameters including shape, pattern, and number of discrete paths needed. Another example is in (Dietz et al. 2016) where the parameters (referred to as "degrees of freedom") of welding a T-joint seam including welding direction and rotation are varied to create unique test cases (see Figure 3).

Measuring Variation and Success

If only a few measurable parameters are varied as part of an evaluation, the "amount" of possible variation can be quantified and presented as a threshold; e.g., +/- 5 cm variation in object placement on a surface. As the number of parameters increases, though, it is likely infeasible to quantify all of them. Instead, qualitative groupings of unique test cases can be generated to articulate the context. Using the example composite layup target objects in Figure 2, the parameters can each be given a set of discrete qualifiers that can be attached to each test case; e.g., shape: concave or convex; pattern: lawnmower or spiral. The metrics derived from these test cases could then be generalized to others with similar qualities (Amigoni, Luperto, and Schiaffonati 2017). A similar "feature-extraction" technique is taken when using the ASTM F3381 standard (ASTM 2019), which identifies several parameters of stationary obstacles that could be encountered by an industrial mobile robot. Qualitative parameters in that standard include shape, material, color, and face quality (e.g., open, closed). The parameters are selected because some combinations are known to be problematic for certain sensors used in obstacle avoidance (e.g., flat black surface with lidar (Kneip et al. 2009)).

It is desirable to record the precise parameters present for every repetition of a task, particularly if performance is shown to be inconsistent under certain conditions. If testing is performed in simulation then recording these parameters should be feasible. In physical testing it may prove difficult to record the conditions of every task repetition if there are multiple parameters being varied and if the speed of operation is fast (e.g., rapid pick and place). Additional sensors can assist in establishing ground truth measurements of the environment, such as by outfitting target objects with IMUs for orientation data, affixing fiducials to track position via cameras, or utilizing additional sensors on the robot. For example, (Morrow et al. 2018) outfitted robotic hands with potentiometers to provide ground truth measurements of joint angles while evaluating grasping.

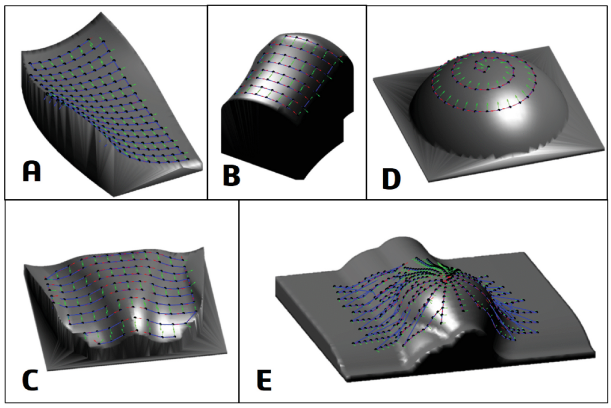

Figure 2: Example target objects for a composite layup task that each require different trajectories to be planned. Adapted from (Malhan et al. 2019) with permission.

For any evaluation of a robot's capabilities, multiple repetitions of performance must be elicited such that statistically significant measures can be obtained. Success criteria should be set for every repetition (e.g., target object must be sanded in 10 minutes or less) and for the test as a whole (e.g., must successfully complete 29 repetitions). The number of successful repetitions required is associated to a probability of success and confidence in the results, ranging from 10 to 459 successful repetitions with no failures for 85% to 99% probability of success (respectively) with 80% to 99% confidence (respectively). To achieve similar probability of success and confidence while allowing for some failures, the number of required successful repetitions increases. See (Leber, Pibida, and Enders 2019) for more detail. These statistics assume that the experiment stimuli used in all repetitions is static. The introduction of variance into some of the stimuli may impact how to perform these types of evaluations.

Inducing Variation

Before inducing variation into an evaluation, a baseline of performance should be established. When varying local contextual characteristics (i.e., environmental perturbations or variations in characteristics of target object or input data), the baseline should be established under ideal—but realistic—conditions. When varying global contextual characteristics (i.e., markedly different environments, tasks, or robots), the baseline can be established from an intended initial scenario. This type of variance is typically induced when demonstrating the versatility of a system to serve multiple, distinctly different functions. The baseline may consist of performance measures derived in the initial context where the robot is planned to be used. Variations may be induced systematically (i.e., selecting from a set of test cases) or stochastically. The latter may be necessary if relying on naturally occurring variance, such as the location and orientation of objects delivered via chute to a platform before they are grasped.

After the baseline is established, another evaluation should be conducted that includes the induced variation.

This is similar to the introduction of "agility challenges" in (Downs, Harrison, and Schlenoff 2016), which compares performance of a static task to a dynamic one that induces interruptions and reprioritizations. During the 2017 Agile Robotics for Industrial Automation Competition (ARIAC) Competition (which used the test methods in the referenced paper), one team was able to place by studying the timing and patterns of when the events were triggered and thus did not require an AI solution (Harrison, Downs, and Schlenoff 2018). Variations like this that are more event-based must consider timing as a parameter to be varied during task repetitions as they otherwise risk being able to be gamed.

Aside from preventing gaming, overfitting of an autonomous solution is also a concern. For example, a mobile robot may encounter an obstacle and add the obstacle to their cost map such that it can plan around that obstacle when the area is traversed again. At this point, the robot is technically no longer detecting and avoiding the obstacle on every subsequent repetition. To counteract this, the characteristics of the obstacle can be varied in between repetitions, intentionally modifying its features and/or position in the opposite part of the aisle that was previously traversed, a concept described in (Norton and Yanco 2016). A similar technique was used in one of the experiments for the DARPA Fast Lightweight Autonomy program wherein UAVs autonomously navigated several test courses in a warehouse. Each test course had a unique overall design (e.g., single corridor, multiple corridors with a turn, etc.), but multiple runs were conducted in each course with different obstacle layouts to test each system's robustness (Mohta et al. 2018).

Robots operating in manufacturing environments are often provided input data that will assist in performing its task, such as a CAD model of a target object for identification purposes or a map for navigation. With this information provided, the robot may select from a set of predefined actions or plan its actions and then modify based on sensed misalignments between expectation and reality. (Gill et al. 2019) induces variation along this parameter into the evaluation of a mobile robot for automotive manufacturing, intentionally misaligning the presence of obstacles physically and/or virtually (e.g., present in the robot's map, but not physically, or in a different location). This parameter has been formalized as "knowledge conditions" in a test method for evaluating navigation and obstacle avoidance capabilities (Norton, Gavriel, and Yanco 2019). Another common variable to input data for manufacturing systems is human interaction to program it to perform a task. Relevant parameters that could be varied for HRI evaluation include user experience and the complexity of a task that must be programmed.

Evaluating Learning Robots

Machine learning is a technique aimed at maintaining and building capability in the presence of variation. AI-enabled robots that are able to learn can do so offline or online. Training an offline learning system is critical to its success and presents another avenue for potential variation; parameters can include the number of training samples, characteristics of the samples used for training, and training on synthetic or physical data. The latter has shown to be problematic

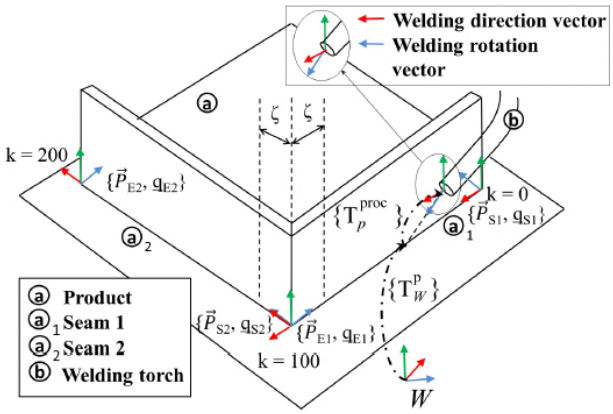

Figure 3: Parameters of welding a T-joint seam that influence required different robot capabilities: welding directions (red arrows) and welding rotations (blue arrows). Adapted from (Dietz et al. 2016) with permission.

as some simulators cannot accurately model certain sensors (e.g., contact forces (Xu et al. 2018)). Learning from demonstration can also be subject to variances inherent in the human teacher. For example, in (Liu and Zhang 2015), training data for a robot learning to weld is obtained from human demonstration wherein the welding current is randomly varied to elicit appropriate response data from the human. The robot is then evaluated in experiments where artificial error is induced systematically to mirror the induced variance during training, albeit in a controlled manner.

Online learning systems will adjust their performance over time as they continue to operate, hopefully improving the system performance. Conducting performance evaluations of these systems is very challenging because of the additional parameters at play (e.g., supervision, reinforcement, amount of learning each task) and the compounding nature of learning. Continuous or lifelong robot learning approaches are typically evaluated against other algorithms or in comparison to a baseline without learning. Another technique is to conduct progressive experiments where a new task is continually revisited as experience in previous tasks grows (Chen and Liu 2016). When reporting metrics derived through this kind of progressive experimentation, the "amount" of experience learned (or at least exposed to) at each testing interval should be quantified and reported alongside the derived metrics, as well as the threshold of variation induced at each stage (as described previously). Some techniques from machine learning algorithms can be leveraged here, too, such as introducing adversarial data sets, inducing expected perturbations to a data set, and mutation testing. See (Braiek and Khomh 2018) for a more comprehensive survey of techniques.

As online learning continues for longer periods of time, systems can forget certain learned techniques. As described in (Farquhar and Gal 2018), it is easier for a continual learning system to learn a completely new task without forgetting than learning a task that is similar to one it has already learned. This presents an opportunity for evaluating

a learning system's memory by varying parameters including the order in which tasks are learned. There is also the concept of forward and backward learning transfer, a metric that can be evaluated by intentionally varying learning order. This and other elements of continual learning are outlined in (Lesort et al. 2019) where recommendations that include evaluation are provided, such as recommending to evaluate algorithms on more than two tasks and using publicly available baselines and benchmarking tools to improve reproducibility. Datasets and benchmarks are available for continuous object recognition such as CORe50 (Lomonaco and Maltoni 2017), but much less for other robotic and manufacturing oriented tasks. Regardless of the benchmarking tools used, conducting evaluations of continuous learning is still in need of more robust approaches to exhaustive benchmarking and evaluation schemes (Parisi et al. 2019).

Evaluating Explainable Robots

An explainable robot system is one that can provide insight to another agent as to its capabilities or why it performed a certain way. Efforts for the latter to develop robots that are able to self-assess their proficiency before, while, or after executing a task seek to tackle this problem (Han et al. 2019; Lee 2019) with the goal of producing more fluent HRI. As learning is introduced for robots in manufacturing and new tasks are learned, explainability may be required throughout the system's life cycle. For an outcomes-based evaluation of these capabilities, relevant metrics will pertain to the communication of an explanation, which can relate to how the communication is performed and how the human agent reacts to the communication. Both axes of evaluation can inform one another.

Success criteria for evaluating an explanation can include whether or not a provided explanation is correct, warranted, and/or understood. Example metrics include qualities of the communication itself (e.g., succinctness, latency, communication modality) and how the human agent interprets the communication (e.g., perceptability, understandability, processing difficulty). Evaluation will involve an experiment design wherein the next action to be taken by a human agent after receiving an explanation is dependent on the accuracy of the robot's explanation and/or the human's interpretation of it. The conceptual model of processing an explanation provided in (Hoffman et al. 2018) outlines this type of evaluation as a "test of comprehension" (i.e., alignment of the user's mental model) followed by a "test of performance." This produces a three-step evaluation process: (1) the robot performs a task, (2) the robot communicates an explanation as to its performance on that task, and (3) the human's performance in reaction to that communication. These steps may happen in a different order depending on when the explanation occurs (e.g., if a priori: 2, 3, 1).

Recommendations

With the previously described considerations in mind, several recommendations are made to advance metrics and evaluation methods for AI-enabled robots in manufacturing.

Develop guidance and resources for deriving salient parameters of manufacturing contexts that can be varied as part of evaluation. One option to support this is the development of taxonomies of relevant robot capabilities and characteristics and correlating the associated challenges posed when used in a manufacturing setting. Common issues associated with different robotic components, tasks, or processes in manufacturing environments should be cataloged. Reference materials like (Boothroyd, Dewhurst, and Knight 2010) that define the various components and characteristics of manufacturing tasks and processes should be leveraged as starting points for development.

Generate more manufacturing-oriented data sets and benchmarks. Evaluation resources like data sets for object recognition (Lomonaco and Maltoni 2017) and benchmarking tools for robotic manipulation (Calli et al. 2015) have proven beneficial for the research community, but are not focused on manufacturing. An effort out of the National Institute of Standards and Technology is to develop data sets and models specifically for AI and machine learning of robots in manufacturing (National Institute of Standards and Technology 2020).

Develop benchmarking tools and test methods for robot learning and explainability. These burgeoning capabilities for manufacturing require careful consideration when being evaluated and can leverage techniques from non-robotic machine learning evaluation and HRI metrics. For example, efforts like REPLAB (Yang et al. 2019)) which provide a common platform, task definition, and protocols for benchmarking robot learning of grasping. There do not appear to be any common resources towards evaluating explainability in robots.

Develop tools to automatically induce variation and record ground truth measurements. Physical methods that can automatically reset the position of target objects in order to conduct evaluation with a high number of required task repetitions to achieve statistical significance can ease the burden of long-term evaluation. Additionally, the ability to measure and record the events of an evaluation through sensorizing the test space can greatly increase fidelity.

Define the scope of an evaluation and classification of metrics using a common lexicon and framework. The context of an evaluation, the metrics used, what element(s) they correspond to, and the comparison of results are important factors when designing evaluations. Similar to the benchmarking scheme for bipedal locomotion (Torricelli et al. 2015), a framework and standard terminology set for robots in manufacturing can be used to guide the design of an evaluation and in communicating these concepts. Following this recommendation, an evaluation framework is proposed in the next section.

Proposed Evaluation Framework

The framework consists of several components that can assist in defining and communicating the scope of an evaluation. See Figure 4. This framework is under development as part of an effort to standardize metrics used in projects funded by the Advanced Robotics for Manufacturing (ARM) Institute (ARM Institute 2020), but aims to be

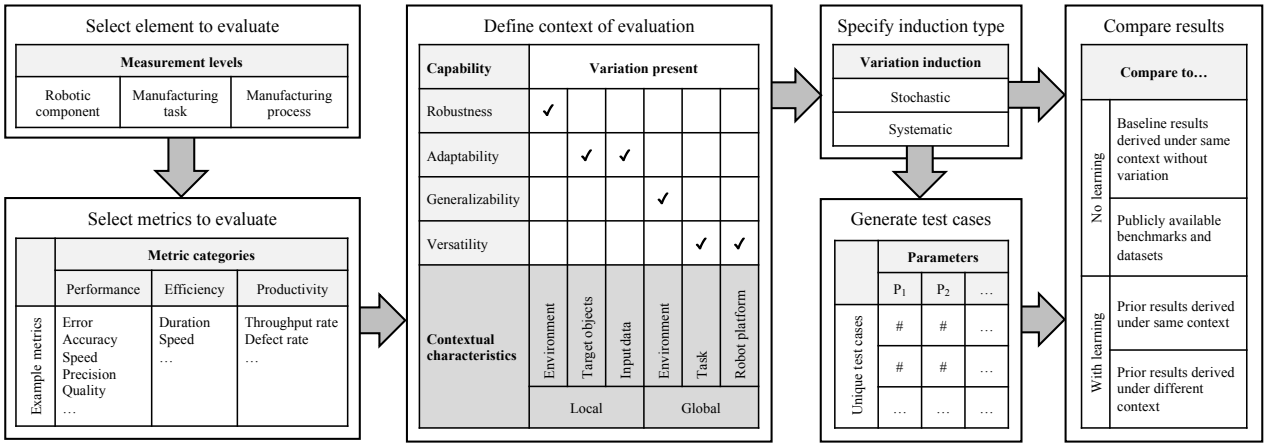

Figure 4: Framework in development for designing evaluations and classifying metrics of AI-enabled robots in manufacturing.

more broadly applicable to other outcomes-based evaluations of robots in manufacturing.

Three *measurement levels* are defined in order to specify what element(s) the metrics used in an evaluation are being used to measure: robotic component, manufacturing task, and manufacturing process. Three *metrics categories* are also defined: performance, efficiency, and productivity. By combining these two components, one can express what the evaluated metric is representative of; e.g., measuring the performance of a planner (robotic component) by measuring the speed at which it calculates trajectories, the efficiency of transferring a grasped object to a bin (manufacturing task) by measuring its duration, the productivity of kitting (manufacturing process) by measuring throughput rate.

The *contextual characteristics* that can be varied as part of an evaluation are divided into local (environment, target objects, input data) and global (environment, task, robot). These are corresponded to four capabilities that an AI-enabled robot can possess, each of which are defined according to what type of variability the robot is able to withstand:

- *Robustness*: local variations in the environment; e.g., fluctuations in ambient lighting or target object location.
- *Adaptability*: local variations in target object and/or input data characteristics; e.g., varied geometry of objects that all require the same robot competencies to be effectively interacted with.
- *Generalizability*: global variations in environment; e.g., for a mobile robot, changing from navigating through open spaces with less features to narrower, confined feature-dense aisles.
- *Versatility*: global variations in task and/or robot platform; e.g., for a trajectory planner of a robot manipulator, changing the task from kitting (coarser object placement) to welding (sensitive adherence to planned paths).

If a robot possesses the capability, then evaluations conducted under defined variation of the related contextual characteristics will have been successful. These terms are commonly used throughout manufacturing with many possible definitions, but no single unified definitions exist. Their definitions in this paper are intended to delineate them from one another purely for evaluation purposes.

If variation is induced *systematically*, then multiple unique test cases will be generated and evaluated. *Stochastic* variation may be less measurable for ever task repetition and may occur naturally as part of a task design (e.g., position of an object to be inspected may slightly shift due to human error in placing it in front of the robot). The results of the evaluation can be compared to baselines, publicly available benchmarks, or to prior results if the robot is capable of learning. The framework can also be used to design experiments in order to elicit appropriate training data for a learning system, and then later used to evaluate that capability under new context.

The terms used in the framework are intended to simplify specification of an evaluation and the interpretation of results. For example, one could evaluate the adaptability of a robot while performing a sanding task by varying the target objects to include those with similar but unique geometry. The performance of the manufacturing task can be evaluated by measuring speed (e.g., time per workpiece) and quality (e.g., surface roughness in microinches) and comparing it to a baseline of a single workpiece. If the robot learns, additional evaluations can be conducted at different intervals after so many workpieces have been sanded for comparison.

The prototype framework does not purport to address all evaluation needs of AI-enabled robots. It is presented as a starting point for cultivating discussion and development of metrics and evaluation methods for these systems and will continue to be refined in the future.

Conclusion

This paper presented several considerations for developing metrics and evaluation methods for AI-enabled robots in manufacturing, including how to select, measure, and induce variation as part of an evaluation, evaluating a robot's ability to learn, and if the robot is explainable. Recommendations were made to advance the development of metrics and eval-

uation methods, outlining gaps related to context definition, benchmarking, automated testing, and guidance for designing evaluations. In response to the latter recommendation, a prototype framework that corresponds several elements of evaluating AI-enabled robots in manufacturing is presented.

There are additional efforts by the authors and collaborators following some of the recommendations made in this paper, such as developing physical testing infrastructure outfitted with sensors for ground truth measurements and automation systems to support systematic evaluations of robots (National Science Foundation 2020), and developing metrics for explainability as part of the SUCCESS MURI project (CMU, BYU, Tufts, UML 2020). The proposed framework is still under development and the authors encourage feedback on its design such that it can be more usable in guiding evaluations. We will continue development such that more concrete evaluation guidelines can be generated as the research field advances.

Acknowledgments

This work has been supported in part by the Advanced Robotics for Manufacturing (ARM) Institute (Award #28158), the Office of Naval Research (N00014-18-1-2503), the National Science Foundation (CNS-1925604), and the National Institute of Standards and Technology (70NANB17H256).

References

Amigoni, F.; Luperto, M.; and Schiaffonati, V. 2017. Toward generalization of experimental results for autonomous robots. *Robotics and Autonomous Systems* 90:4–14.

ARM Institute. 2020. Advanced robotics for manufacturing. http://www.arminstitute.org/. Accessed: 2019-01-14.

ASTM. 2019. ASTM F3381-19, Standard Practice for Describing Stationary Obstacles Utilized within A-UGV Test Methods, ASTM International, West Conshohocken, PA, 2019, www.astm.org.

Boothroyd, G.; Dewhurst, P.; and Knight, W. A. 2010. *Product Design for Manufacture and Assembly*. CRC Press.

Bostelman, R.; Foufou, S.; Legowik, S.; and Hong, T. H. 2016. Mobile manipulator performance measurement towards manufacturing assembly tasks. In *IFIP International Conference on Product Lifecycle Management*, 411–420. Springer.

Braiek, H. B., and Khomh, F. 2018. On testing machine learning programs. *arXiv preprint arXiv:1812.02257*.

Browne, J.; Dubois, D.; Rathmill, K.; Sethi, S. P.; Stecke, K. E.; et al. 1984. Classification of flexible manufacturing systems. *The FMS magazine* 2(2):114–117.

Calli, B.; Singh, A.; Walsman, A.; Srinivasa, S.; Abbeel, P.; and Dollar, A. M. 2015. The ycb object and model set: Towards common benchmarks for manipulation research. In *2015 international conference on advanced robotics (ICAR)*, 510–517. IEEE.

Ceballos, N. D. M.; Valencia, J. A.; and Ospina, N. L. 2010. *Quantitative performance metrics for mobile robots navigation*. INTECH Open Access Publisher.

Chen, Z., and Liu, B. 2016. Lifelong machine learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 10(3):1–145.

CMU, BYU, Tufts, UML. 2020. SUCCESS MURI - Self-assessment and Understanding of Competence and Conditions to Ensure System Success. https://successmuri.org/. Accessed: 2019-01-15.

Dietz, T.; Ockert, P.; Kuss, A.; Hägele, M.; Verl, A.; et al. 2016. Automatic optimal motion generation for robotic manufacturing processes: Optimal collision avoidance in robotic welding. In *2016 IEEE International Conference on Automation Science and Engineering (CASE)*, 154–161. IEEE.

Downs, A.; Harrison, W.; and Schlenoff, C. 2016. Test methods for robot agility in manufacturing. *Industrial Robot: An International Journal* 43(5):563–572.

ElMaraghy, H. A. 2005. Flexible and reconfigurable manufacturing systems paradigms. *International journal of flexible manufacturing systems* 17(4):261–276.

Falco, J.; Hemphill, D.; Kimble, K.; Messina, E.; Norton, A.; Ropelato, R.; and Yanco, H. 2020a. Benchmarking protocols for evaluating grasp strength, grasp cycle time, finger strength, and finger repeatability of robot end-effectors. *IEEE Robotics and Automation Letters* 1–1.

Falco, J.; Kimble, K.; Van Wyk, K.; Messina, E.; Sun, Y.; Shibata, M.; Uemura, W.; and Yokokohji, Y. 2020b. Benchmarking protocols for evaluating small parts robotic assembly systems. *IEEE Robotics and Automation Letters* 1–1.

Falco, J. A.; Van Wyk, K.; and Messina, E. R. 2018. Performance metrics and test methods for robotic hands (draft). Technical report.

Farquhar, S., and Gal, Y. 2018. Towards robust evaluations of continual learning. *arXiv preprint arXiv:1805.09733*.

Gill, J. S.; Tomaszewski, M.; Jia, Y.; Pisu, P.; and Krovi, V. N. 2019. Evaluation of navigation in mobile robots for long-term autonomy in automotive manufacturing environments. Technical report, SAE Technical Paper.

Han, Z.; Allspaw, J.; Norton, A.; and Yanco, H. A. 2019. Towards a robot explanation system: A survey and our approach to state summarization, storage and querying, and human interface. *arXiv preprint arXiv:1909.06418*.

Harrison, W.; Downs, A.; and Schlenoff, C. 2018. The agile robotics for industrial automation competition. *AI Magazine* 39(4):77.

Hoffman, R. R.; Mueller, S. T.; Klein, G.; and Litman, J. 2018. Metrics for explainable ai: Challenges and prospects. *arXiv preprint arXiv:1812.04608*.

Kneip, L.; Tâche, F.; Caprari, G.; and Siegwart, R. 2009. Characterization of the compact hokuyo urg-04lx 2d laser range scanner. In *2009 IEEE International Conference on Robotics and Automation*, 1447–1454. IEEE.

Leber, D. D.; Pibida, L. S.; and Enders, A. L. 2019. Confirming a performance threshold with a binary experimental response. Technical report.

Lee, Y. T.; Kumaraguru, S.; Jain, S.; Robinson, S.; Helu, M.; Hatim, Q. Y.; Rachuri, S.; Dornfeld, D.; Saldana, C. J.; and

Kumara, S. 2017. A classification scheme for smart manufacturing systems' performance metrics. *Smart and sustainable manufacturing systems* 1(1):52.

Lee, M. S. 2019. Self-assessing and communicating manipulation proficiency through active uncertainty characterization. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 724–726. IEEE.

Lesort, T.; Lomonaco, V.; Stoian, A.; Maltoni, D.; Filliat, D.; and Díaz-Rodríguez, N. 2019. Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges. *Information Fusion*.

Liu, Y.-K., and Zhang, Y.-M. 2015. Supervised learning of human welder behaviors for intelligent robotic welding. *IEEE Transactions on Automation Science and Engineering* 14(3):1532–1541.

Lomonaco, V., and Maltoni, D. 2017. Core50: a new dataset and benchmark for continuous object recognition. *arXiv preprint arXiv:1705.03550*.

Mahler, J.; Platt, R.; Rodriguez, A.; Ciocarlie, M.; Dollar, A.; Detry, R.; Roa, M. A.; Yanco, H.; Norton, A.; Falco, J.; et al. 2018. Guest editorial open discussion of robot grasping benchmarks, protocols, and metrics. *IEEE Transactions on Automation Science and Engineering* 15(4):1440–1442.

Malhan, R. K.; Kabir, A. M.; Shah, B.; and Gupta, S. K. 2019. Identifying feasible workpiece placement with respect to redundant manipulator for complex manufacturing tasks. In *2019 International Conference on Robotics and Automation (ICRA)*, 5585–5591. IEEE.

Marvel, J. A.; Saidi, K.; Eastman, R.; Hong, T.; Cheok, G.; and Messina, E. 2012. Technology readiness levels for randomized bin picking. In *Proceedings of the Workshop on Performance Metrics for Intelligent Systems*, 109–113. ACM.

Mohta, K.; Watterson, M.; Mulgaonkar, Y.; Liu, S.; Qu, C.; Makineni, A.; Saulnier, K.; Sun, K.; Zhu, A.; Delmerico, J.; et al. 2018. Fast, autonomous flight in gps-denied and cluttered environments. *Journal of Field Robotics* 35(1):101–120.

Morrow, J.; Kothari, A.; Ong, Y. H.; Harlan, N.; Balasubramanian, R.; and Grimm, C. 2018. Using human studies to analyze capabilities of underactuated and compliant hands in manipulation tasks. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2949–2954. IEEE.

National Institute of Standards and Technology. 2020. Embodied ai and data generation for manufacturing robotics. https://www.nist.gov/programs-projects/embodied-ai-and-data-generation-manufacturing-robotics. Accessed: 2019-01-11.

National Science Foundation. 2020. CCRI: Medium: Collaborative Research: Physical Robotic Manipulation Test Facility. https://www.nsf.gov/awardsearch/showAward?AWD_ID=1925715 and https://www.nsf.gov/awardsearch/showAward?AWD_ID=1925604. Accessed: 2019-01-16.

Norton, A., and Yanco, H. 2016. Preliminary development of a test method for obstacle detection and avoidance in industrial environments. In *Autonomous Industrial Vehicles: From the Laboratory to the Factory Floor*. ASTM International.

Norton, A.; Gavriel, P.; and Yanco, H. 2019. A standard test method for evaluating navigation and obstacle avoidance capabilities of agvs and amrs. *ASTM Journal of Smart and Sustainable Manufacturing Systems* 3(2):106–126.

Norton, A.; Messina, E.; and Yanco, H. 2020, In Press. Advancing capabilities of industrial robots through evaluation, benchmarking, and characterization. *Recent Advances in Industrial Robotics*.

Ortenzi, V.; Controzzi, M.; Cini, F.; Leitner, J.; Bianchi, M.; Roa, M.; and Corke, P. 2019. Robotic manipulation and the role of the task in the metric of success. *Nature Machine Intelligence* 1(8):340–346.

Parisi, G. I.; Kemker, R.; Part, J. L.; Kanan, C.; and Wermter, S. 2019. Continual lifelong learning with neural networks: A review. *Neural Networks*.

Shneier, M. O.; Messina, E. R.; Schlenoff, C. I.; Proctor, F. M.; Kramer, T. R.; and Falco, J. A. 2015. Measuring and representing the performance of manufacturing assembly robots. Technical report.

Steinfeld, A.; Fong, T.; Kaber, D.; Lewis, M.; Scholtz, J.; Schultz, A.; and Goodrich, M. 2006. Common metrics for human-robot interaction. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, 33–40. ACM.

Torricelli, D.; Gonzalez-Vargas, J.; Veneman, J. F.; Mombaur, K.; Tsagarakis, N.; del Ama, A. J.; Gil-Agudo, A.; Moreno, J. C.; and Pons, J. L. 2015. Benchmarking bipedal locomotion: a unified scheme for humanoids, wearable robots, and humans. *IEEE Robotics & Automation Magazine* 22(3):103–115.

Xu, J.; Hou, Z.; Wang, W.; Xu, B.; Zhang, K.; and Chen, K. 2018. Feedback deep deterministic policy gradient with fuzzy reward for robotic multiple peg-in-hole assembly tasks. *IEEE Transactions on Industrial Informatics* 15(3):1658–1667.

Yang, B.; Zhang, J.; Pong, V.; Levine, S.; and Jayaraman, D. 2019. Replab: A reproducible low-cost arm benchmark platform for robotic learning. *arXiv preprint arXiv:1905.07447*.